基于生命周期模型的科技文献数据管理体系研究

常志军[1,2]，许丽媛[1*]，于倩倩[1]，张建勇[1,2]，王永吉[3]

(1. 中国科学院文献情报中心 数据资源部，北京 100190；2. 中国科学院大学 图书情报与档案管理系，北京 100049；3. 中国科学院软件研究所 计算机科学国家重点实验室，北京 100190)

摘　要：[目的/意义]科技文献数据资源具有覆盖广、数量大、类型多、更新快、时效强等特点，为提高科技文献数据管理效果和数据安全，本文基于数据生命周期模型对科技文献管理体系进行研究。[方法/过程]对科技文献管理模式进行探索，基于数据管理流程，构建了科技文献的生命周期体系，并从数据创建、数据存储、数据预处理、数据计算、数据服务、数据归档、数据销毁等7个阶段对数据管理工具和数据管理方法进行阐述。[结果/结论]本文对科瑞唯安核心数据集 WOS BP 数据进行了基于科技文献生命周期的管理和实践，同时基于 DAMA 数据质量的 6 个评估维度对数据管理效果进行综合评价。

关键词：生命周期管理；科技文献；数据管理；大数据治理；知识图谱

中图分类号：TN919；G250　　文献标识码：A　　文章编号：1002-1248（2022）06-0036-14

引用本文：常志军,许丽媛,于倩倩,等.基于生命周期模型的科技文献数据管理体系研究[J].农业图书情报学报,2022,34(6):36-49.

1 引　言

数据是一种重要的资产[1]，与事物资产的可见可动，财务资产的可计可量不同，数据资产有其独特的特性和价值：持久保存性、损坏不再生、使用无消耗、动态应用性、多状态应用、数据自生产等。数据管理的核心是确保数据的质量，如果数据未能满足使用者的需求，那么所有收集、存储、安全加固、使用数据的努力都是无用的。据 IBM 估算，2016 年，美国由于数据质量问题而导致的损失达到 3.1 万亿美元[2]。因此，数据使用者必须与具备专业知识领域和技能的数据管理团队共同参与定义数据的特征，使之成为高质量的数据。

中国科学院文献情报中心（以下简称文献中心）通过集团采购、资源置换、自主建设等渠道收集了大量的科技文献，包括科技图书、科技期刊、科技报告、专利文献、会议文献、学位论文、标准文献等。这些科技文献类型众多，来源广泛，凝聚着人类在科技探索过程中的经验和智慧[3]。在信息化迅速发展的当下，

收稿日期：2022-03-01

基金项目：国家社会科学基金“面向循证医学的领域文献实体关系识别方法研究”（21BTQ106）

作者简介：常志军，男，硕士，副研究馆员，硕士研究生导师，研究方向为大数据平台建设与管理、领域知识图谱构建等。于倩倩，女，硕士，副研究馆员，研究方向为数据管理与组织。张建勇，男，硕士，研究馆员，研究方向为数据管理与组织。王永吉，男，博士，研究员，研究方向为计算机科学与技软件工程等

*通信作者：许丽媛，女，硕士，馆员，研究方向为数据资源管理与质量控制。Email：xuly@mail.las.ac.cn

如何对这些科技文献进行有效、高效的管理是亟需面对和解决的问题，也是文献中心科技文献管理工作发展的重要方向。

国内外很多研究团队开展了科技文献管理方法的研究和科技文献管理体系的建设，如云安全联盟组织为云环境数据提出 CSA 模型[4]，包括创建、存储、使用、共享、存档和销毁，他是为云环境设计的，重点解决了数据安全，未考虑数据质量、数据处理和数据分析等内容。美国地质调查局数据集成社区提出采用 USGS 模型管理数据，包括计划、获取、处理、分析、保存和发布 / 共享，用于评估和改进管理科学数据的政策和实践，是一个综合的模型[5,6]。大学间政治和社会研究联合会提出采用 DDI 模型[7]管理数据，包括研究概念、数据收集、数据处理、数据存档、数据分发、数据发现、数据分析和重新调整用途等，是一个全面的模型，但是没有对数据质量和数据安全的关注。张迎等[8]提出了科学数据管理生命周期，并从获取、描述、存储、发布、重用等 5 个阶段对科学数据进行管理。

但当前就如何利用生命周期理论对科技文献进行综合管理和有效利用，以及采用专业的衡量标准进行质量评估等研究还处在初级阶段。围绕基于生命周期理论对科技文献进行综合管理等需求，本文第二部分论述了数据生命周期管理模型，总结归纳符合科技文献生命周期发展的阶段和模型，本文第三部分重点介绍了数据管理体系研究的 7 个流程，并详细说明了每个阶段的管理体系建设内容，本文第四部分创新性的以 WOS BP 数据为基础开展基于生命周期的数据管理实践，并依照数据管理目标从 6 个维度进行管理实践与综合评价，本文第五部分对工作进行简要总结，并对未来工作进行展望。

2 数据生命周期管理

2.1 数据生命周期管理模型

数据不是静止的，在整个生命周期中，数据需要被清洗、转换、合并、增强等。不同类型的数据具有不同的生命周期，这加大了数据生命周期中相关概念的复杂性。如事务型数据可以通过基本业务规则得到管理，而主数据需要通过数据综合处理得到管理。尽管如此，仍然存在一些生命周期通用规则，适用于任何数据。2018 年国务院办公厅在印发《科学数据管理办法》[9,10]时指出要加强科学数据全生命周期管理[11]，规范科学数据的采集生产、加工整理、开放共享等各个环节的工作。同时也将科学数据管理生命周期分为数据采集和交汇、数据保存、数据共享利用、数据保密安全等方面。

数据生命周期管理（Data Life Cycle Management，DLM）是一种基于策略的方法[12]，着重于数据的规划和设计、使数据可用、可维护，以及通过应用数据实现组织的目标，最终达到可被需要的人或流程所使用的目的。通常用于管理数据在整个生命周期内的流动：从数据的创建和初始存储、变化、迁移和维护到它过时被删除的全过程[13]。尽管数据和技术是交织在一起的，但是不能把数据的生命周期混淆为系统开发生命周期（Systems Development Life Cycle，SDLC），因为系统开发生命周期专注于在预算范围内按时完成项目研发任务[14]。

数据生命周期管理模型定义从生产阶段到服务阶段的数据全景视图，目标是优化数据管理、提高效率、降低成本。DAMA 数据资产管理协会作为一个全球性的数据管理协会，致力于数据管理的研究和实践原则。DAMA 模型包括创建或获取数据、移动、转换和存储数据并使其得以维护和共享的过程、使用数据的过程以及处理数据的过程[15]。在数据的整个生命周期中，可以清理、转换、合并、增强或聚合数据，同时随着数据的使用或增强，通常会生成新的数据，因此生命周期具有内部迭代。

基于生命周期管理的数据可以在一定程度上提升数据质量，最终达到数据使用者的期望并满足数据需求。判断数据质量优劣的标准是与能否满足数据消费者的需求一致为基准，一致则属于高质量数据，反之，不适用于数据使用者的数据则是低质量数据。数据质量维度是数据的可测量特性或属性，为了评估数据的

质量，需要建立具体可行的衡量维度，这些维度不但对业务流程很重要，而且具备可测量、可操作的特性。2013 年，DAMA 英国分会编写的数据管理白皮书提出了 6 个核心的数据质量评估维度[16]，分别是：完整性（Completeness），是评估已存储数据占应存储数据的百分比。唯一性（Uniqueness），是评估任何实体的记录会不会出现多次。实时性（Timeliness），是评估数据体现特定时间点的真实程度。有效性（Validity），是评估数据是否符合相关定义（格式、种类、范围）。准确性（Accuracy），是评估数据描述真实世界对象或事件的精确度。一致性（Consistency），是评估多处对同一个事物的描述不存在差异。

2.2 科技文献生命周期研究

DAMA 表示数据管理是基于数据生命周期的管理，不同类型的数据有不同的生命周期特征。科技文献数据[17]具备数据量大、文件类型多、获取方式和格式多样、更新频率快、时效性强等特点，以文件类型多为例：科技文献通常覆盖期刊、会议录、专著、丛书、文集汇编、工具书、课程、研究论文、专著章节、科技报告、学位论文、课件等多个类型。此外，科技文献数据可描述内容的颗粒度更细化，如 JATS 数据标准包含了 250 余个元素和 130 余个元素属性，NSTL 统一文献元数据标准包含 97 个描述性元素、53 个辅助性元素和 49 个属性[18]。同时，科技文献数据组织模块化加强，通过对细粒度元素的组合形成相对独立又相互关联的实体模块，如期刊、论文、会议、基金、贡献者、机构等多个实体模块。

本文将科技文献数据的全生命周期阶段主要归纳为创建、存储、预处理、计算、服务、归档、销毁等 7 个阶段，可以在科技文献中进行普适性应用。如图 1 所示，数据在每个阶段呈现不同的活跃度，在数据计算阶段和数据服务阶段达到峰值，在数据销毁阶段达到谷值。

数据创建阶段收集从多个来源获取的商业采购数据、开放获取数据、中心自建数据和交换获取数据等，通过网络接口获取、公开网页采集、数据库直接导入、硬件批量拷贝、网络集中下载等多种接入形式，获取期刊论文、会议论文、科技报告、科技专利、基金项目、科技资讯、图书专著、科技政策等各个类型数据。

数据存储阶段针对不同体量、结构的数据进行个性化存储设计。对无需复杂操作的小体量数据采用本地文件系统存储形式，利用单台服务器满足对文件数据、源数据、中间数据的存储需求。对无需复杂操作的大体量数据采用分布式存储形式，利用多台服务器满足对大文件数据的存储需求。对需要复杂操作的结构化数据采用数据库存储形式，对常规业务数据、监测日志数据等进行存储。

数据预处理阶段可以从字段抽取、信息转换、数

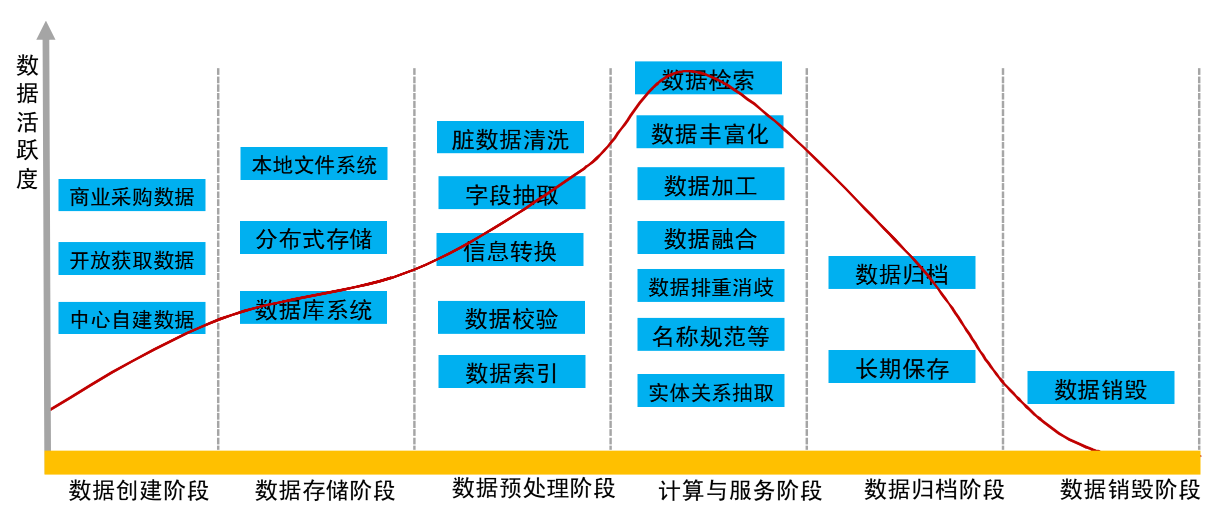


图 1 科技文献生命周期模型

Fig.1 Life cycle model of scientific and technical literature

据清洗、数据校验、数据索引等 5 个方面进行。将非结构化数据转化为符合统一标准的结构化数据，或者将一种形式的结构化数据转化为另一种形式的结构化数据，对相关字段进行抽取、清洗、加工，以获取更丰富更标准的数据，基于服务需求设定索引数据，为了后续计算、服务更方便、灵活。

数据计算阶段是科技文献在整个生命周期中最活跃的阶段，也是数据最具操作性、最丰富的阶段，主要是从数据加工、数据丰富化、数据融合、实体和关系抽取等 4 个方面展开，对数据进行集中的治理和计算，以产生更具使用价值的科研数据。

数据服务阶段是将前期已经处理和计算生成的数据通过各类服务形式稳定、高效地进行数据交互并输出数据，提供基础数据服务和增值数据服务，同时支持面向用户需求定制开发优质的、高效的数据服务，基于权限控制和访问监控保障数据服务安全。

数据归档阶段是将不再经常使用的数据迁移到一个单独的存储设备来进行长期、有效保存的过程，这类数据通常是由旧的数据组成，但又是以后必须参考且很重要的数据，需要长期存储和长期可获取，因此在归档时必须遵从相应的规则进行。

数据销毁阶段是指数据服务到期后进行销毁的过程，通常采用对数据及数据的存储介质物理删除的操作手段，使数据彻底丢失且无法恢复。为保证后续审计需要，在销毁时需要对销毁内容、时间、方式、核准部门及人员等信息进行登记审核。

3 数据管理体系研究

3.1 数据管理流程

基于生命周期进行数据管理的流程主要分为：数据创建登记、数据解析存储、数据加工处理、数据集成计算、数据服务应用、数据归档保存、数据销毁记录等，如图 2 所示。

3.2 数据管理体系

3.2.1 数据创建阶段

在数据创建阶段主要进行数据创建和登记，科技文献数据来源主要分为 3 种类型：商业采购数据、开放获取数据、内部自建数据。各数据来源提供不同的数据获取方式，有些方式利于形成机器自动化例行服务，有些方式需要人工操作获取数据，有些方式利于频繁地、轻量化的获取数据，有些方式则适用于大量数据的快速传递。每一种来源都有其独特的数据接入形式，如表 1 所示。



图 2　数据管理流程

Fig.2　Data management processes

表 1　科技文献数据来源和接入形式

Table 1 Sources and access forms of scientific and technical literature data

数据来源	来源描述	数据类型	数据接入形式
商业采购数据	通过商业购买获取的商业数据资源	期刊论文；会议论文；学位论文；科技报告；科技专利等	OAI 协议下载；数据库导入；移动存储介质拷贝；FTP、SCP、网盘、邮件等形式下载等
开放获取数据	通过各类工具和形式获取的公开数据	开放论文；基金项目；科技政策；科技资讯等	公开数据接口获取；网络页面采集；网盘文件、邮件等形式下载；Github 等公开语料库数据下载等
内部自建数据	通过中心自主构建或与其他团队交换获取的数据资源	业务专用数据；服务日志数据；基于数据产生的处理数据等	数据库导入；移动存储介质拷贝等

不同来源的数据，接入形式各异，因此需要个性化定制多种工具，以满足在数据创建阶段对数据资产的完整接入，如表 2 所示，从接口获取、数据库导入、存储介质拷贝、网络下载、网络采集等方面对数据创建工具进行设计。

3.2.2　数据存储阶段

在数据存储阶段主要进行数据解析和存储，通过各类接入形式获取的科技文献数据主要是 XML、数据表、JSON、文本文件等 4 种类型，对比这几种类型数据主要有以下特点，如表 3 所示。

综合分析科技文献数据的各种数据格式特点，设计统一的文献元数据存储体系，对各类型数据进行统一存储，有助于处理、维护、集成、包含、审计和管理科技文献数据。文献元数据存储体系重点描述了数据本身，如数据库、数据元素、数据模型；数据所代表的概念，如业务流程、应用系统、软件代码、技术基础设施；数据和概念之间的连接和关系等，主要包含了业务元数据、技术元数据和操作元数据 3 类，如表 4 所示。

经过统一文献元数据存储体系描述的科技文献数据资源可以更好的解释、组织、理解各类型数据结构、数据内容、系统业务流程等。以业务元数据为例，根据各来源数据组织结构的特点，为每类实体设计独立存储结构，数据组织、字段命名符合 JATS 数据标准、NSTL 统一文献元数据标准等相关规范，如图 3 所示为科技论文元数据结构。

3.2.3　数据预处理阶段

在数据预处理阶段主要进行数据加工和处理，建

表 2　数据创建工具

Table 2 Data creation tools

数据接入形式	数据创建工具	数据接入形式特点
各类接口获取	建立数据协议工具，基于传递过程中的格式、参数、标准等完成明确的定义	最常用的数据接入形式
数据库导入	建立数据解析工具，基于数据库中的数据项和数据描述进行实质定义	旧数据迁移的常用形式
移动存储介质拷贝	建立数据传递工具，基于移动硬盘、光盘等媒介进行数据传递	适用于接入频率低，单次数据量较大的数据源。需要人工参与操作，无法实现自动获取
FTP、SCP、网盘、邮件等形式下载	建立数据下载工具，基于传统的数据传输协议下载数据	适用于一次性获取或异常状态时临时替代传输形式，数据量大时效率很高。但传输过程中文件可能会损坏，缺少校验机制
网络页面采集	建立数据采集工具，基于网络页面的公开数据进行定向采集	适用于稳定页面的数据采集，页面内容如有变动需要重新定制采集工具
Github 等公开语料库数据下载	建立数据同步工具，基于公开语料库按需获取数据	适用于专业特色数据获取

表 3　数据格式和特点

Table 3 Data formats and features

数据格式	特点	缺点
XML	使用较广泛的文件类型，有通用的格式定义，利于数据传输过程中的校验	各数据来源的 XML 格式各不相同，需要为每种数据源进行详细配置
数据库	字段定义清除明确，便于字段映射	二维结构难以描述字段之间的关联关系，多值处理较为繁琐。数据量大时效率较低
JSON	数据结构灵活，操作方便，传输效率高，软件环境兼容性好	缺少错误处理机制，安全性差
文本文件	最轻便的文件类型，没有额外的编码需求或格式需求	缺少格式定义导致描述形式五花八门，解析映射极为困难。缺少格式校验导致错误难以被发现。已很少使用

表 4　科技文献元数据存储体系

Table 4 Metadata storage system for scientific and technical literature

元数据类型	描述范围	示例
业务元数据	数据内容和状态，与数据治理相关的细节	概念，主题域，实体和属性等非技术性的名称和定义；属性类型和其它属性特征；范围的描述；计算规则；算法和业务规则；有效的阈值范围等
技术元数据	数据技术细节，数据迁移的过程信息	存储数据的系统等
操作元数据	数据处理和访问的详细信息	报告和查询访问模式，频率和执行时间；备份、保留、创建日期、灾难恢复的相关规定等


```
⊟{
    "access_ext-link":"[[[],[[]],[[]],[[\"http://dx.doi.org/10.1016/S0040-6090(96)089
    "access_ext-link-display":"[[[\"http://dx.doi.org/10.1016/S0040-6090(96)08956-0\"]
    "access_ext-link-display_ik":"[[[\"http://dx.doi.org/10.1016/S0040-6090(96)08956-0
    "access_ext-link_ik":"[[[],[[]],[[]],[[\"http://dx.doi.org/10.1016/S0040-6090(96)
    "article_abstract":"[[[\"SnO2 thin films were prepared by the sol-gel process usin
(4)2C(2)H(5)OH as the precursor. After subsequent annealing, the temperature of which
ised silicon, glass or ceramic glass), the films were characterised by scanning elect
er electron spectroscopy. The sensitivity of these films to variations in humidity wa
93% relative humidity.\"]]]",
    "article_abstract_ik":"[[[\"SnO2 thin films were prepared by the sol-gel process u
(4)2C(2)H(5)OH as the precursor. After subsequent annealing, the temperature of which
ised silicon, glass or ceramic glass), the films were characterised by scanning elect
er electron spectroscopy. The sensitivity of these films to variations in humidity wa
93% relative humidity.\"]]]",
    "article_article-id":"[[[\"WOS:A1997WK35900048\"]],[[\"WK359\"]]]",
    "article_article-id-special":"[[[\"WOSBP:WOS:A1997WK35900048\"]],[[\"WOSArchive:WO
    "article_article-id-special_ik":"[[[\"WOSBP:WOS:A1997WK35900048\"]],[[\"WOSArchive
    "article_article-title":"[[[\"sno2thinfilmspreparedbythesolgelprocess\"]]]",
    "article_article-title-en":"[[[\"SnO 2 thin films prepared by the sol-gel process\
    "article_article-title-en_ik":"[[[\"SnO 2 thin films prepared by the sol-gel proce
    "article_article-title_display":"[[[\"SnO2 thin films prepared by the sol-gel proc
    "article_article-title_display_ik":"[[[\"SnO2 thin films prepared by the sol-gel p
    "article_article-title_ik":"[[[\"sno2thinfilmspreparedbythesolgelprocess\"]]]",
    "article_article-type":"[[[\"Article\"]]]",
    "article_article-type-display":"[[[\"期刊论文\"]]]",
    "article_article-type-display_ik":"[[[\"期刊论文\"]]]",
    "article_article-type_ik":"[[[\"Article\"]]]",
    "article_date":"[[[\"1997-01-5\"]]]",
    "article_date_ik":"[[[\"1997-01-5\"]]]",
    "article_harvest-language_bg":"[[[\"E\"]]]",
```


图 3　科技论文元数据结构样例

Fig.3 Example of a technical paper metadata structure

设数据预处理工具，实现对各类型数据的格式预处理、解析、转换、结构化，并存储到目标存储系统，如图 4 所示。

首先，根据数据来源、数据量、接收方式、接收频率的不同，通过简单配置数据解析规则，归纳高复用的数据解析模块，设计基于 HTML、CSV、XML 和 JOSN 等 4 套主要格式的数据解析引擎，形成一套半自动的数据结构化解析处理机制，实现对数据资源的自助收割兼具批量运行的数据组件，为数据深加工做好支撑工作。

然后，对多来源数据进行解析、规范化，生成符合元数据标准格式的数据仓库，同时构建镜像索引，为数据计算提供离线、在线的读取基础。同时完成定时功能实现部分数据源的自动更新。

最后，对汇集的具体字段如学者、机构、关键词、来源等内容进行规范化处理，保证从各数据源采集来的数据可以进行统一的清洗、规范、管理和使用。同时不断完善清洗规则、清洗库，清理冗余字段，提升数据质量，为应用服务提供有效的数据支撑。

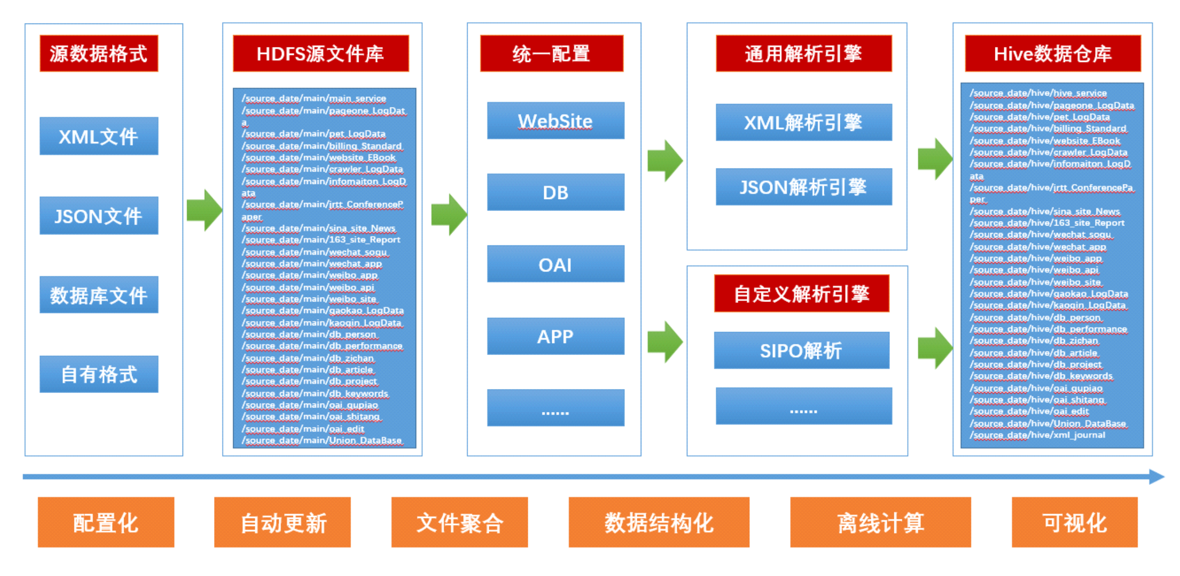


图 4　数据预处理流程

Fig.4　Process of data pre-processing

3.2.4　数据计算阶段

在数据计算阶段主要进行数据集成和计算，主要涉及的关键步骤包括数据丰富化加工、实体关系抽取和知识图谱构建等。

数据丰富化是基于数据已有特征进行信息扩展，提升数据信息量。例如基于文献元数据进行增强关键词扩展，基于摘要进行知识元扩展，基于内容进行中图分类法扩展等。数据加工通常是人工参与的数据加工工作，是最常见的数据质量提升途径。加工过程一般分为加工编辑和审核两个阶段，具有较高的数据质量保障。数据融合是对同一数据的多源处理策略，通常采用优先级筛选和优先占位策略，对不同来源不同类型的数据确定优先等级，质量越高的数据优先级越高，融合时使用来源等级更高的数据字段覆盖来源等级低的字段。当数据字段不能独立支撑数据融合时，可以采取信息块的模式进行综合融合，如图 5 为数据融合流程设计。

数据中往往记录了多个维度或实体的信息，实体

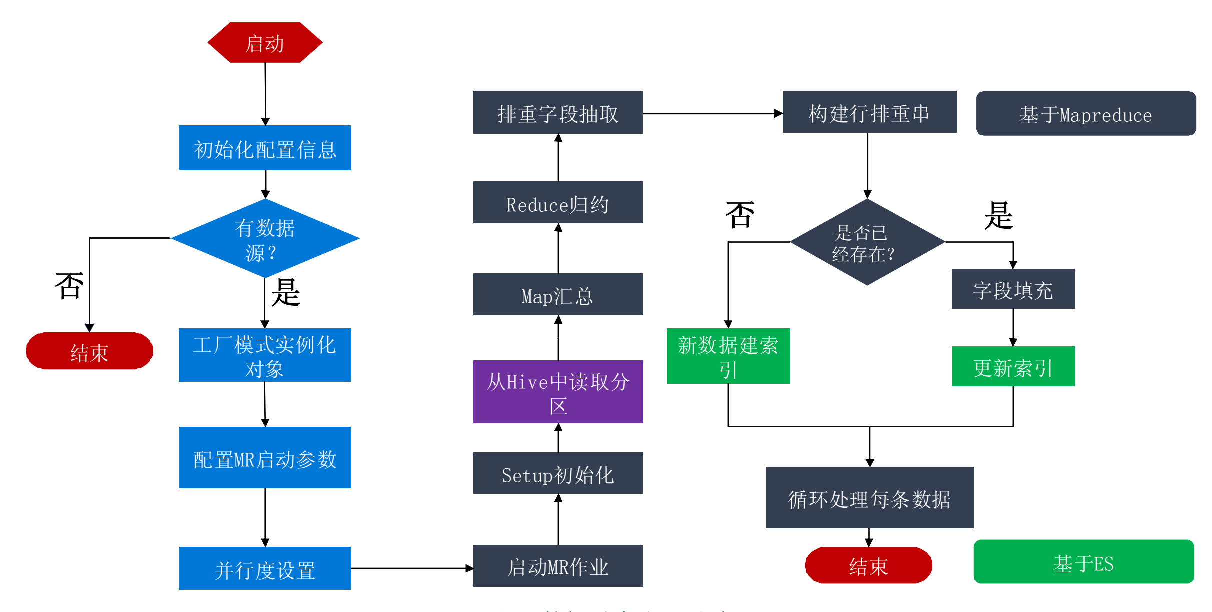


图 5　数据融合流程设计

Fig.5　Design of data fusion process

抽取是基于预处理和丰富化加工后的数据，根据应用需求或知识图谱的设计，定义科研实体和实体间的多维关系模型，从科技文献元数据中提取结构化知识、显化数据间的关联关系、挖掘更深层次的数据内涵、构建学术知识网络关系，形成数据知识图谱，支撑科技文献数据间的关联信息揭示，支持智能知识服务能力。比如从一篇文献中抽取的多个作者实体，隐含着合作作者的关系。

3.2.5　数据服务阶段

在数据服务阶段主要进行数据服务和应用，数据检索是数据服务的主要形式之一，是将数据价值显化的重要手段。采用 SpringCloud 分布式技术体系，设计基于 Eureka、Ribbon、Security、Springboot 等组件的微服务架构，通过 Restful API 接口实现对应用的支撑。微服务技术具有扩展灵活、部署方便、自动负载均衡等特点，以集群模式为多业务提供强稳定、高性能、低延迟的数据服务。如图 6 是数据服务架构。

首先，构建多节点数据注册总线，实现动态服务代理，提供总线基础管理：查看总线使用状态接口，配置安全、注册、监控等功能，通过发布订阅通信应用程序共享信息，通过核心的消息系统负责连接端点和他们之间路由器，以实现数据总线的合理配置。

其次，构建基础设施管理群和服务提供群，部署登录服务、配置服务、查询服务等多个应用服务，可共享数据通路，也可独立部署使用。

最后，构建业务服务消费群，部署数据监控服务，数据分析服务、用户画像服务、检索系统服务等，同时支持各类业务服务的灵活扩展，只需要遵循协议对接到数据总线即可。用户根据需求和应用类型选择适当的接口，通过简单配置 IP、数据通路、offset 等信息，即可通过总线轻松获取数据。

3.2.6　数据归档阶段

在数据归档阶段主要进行数据的归档和保存，在大数据成为了关键资源的今天，归档各种类型的数据是非常重要的，是数据量和数据体量积累的重要阶段。在数据归档时既要考虑存储海量数据的设备成本，也要考虑存储海量数据的时间成本。

基于整个数据生命周期，制定符合业务需求的数据归档策略。首先是识别哪些数据应该被归档，以及需要被归档多长时间。其次，根据数据特性将数据存储在相应的存储设备上，始终将归档数据保留在高性能存储平台上，会导致不必要的成本和人力资源的消

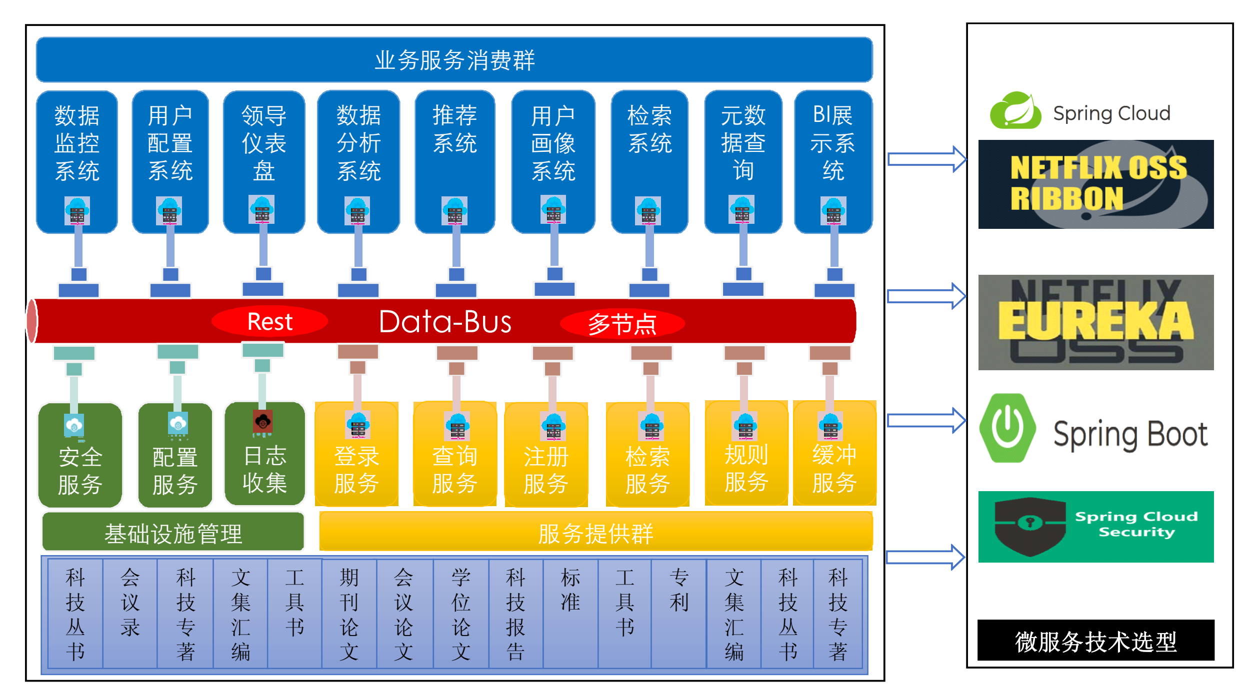


图 6　数据服务架构设计

Fig.6　Design of data service architecture

耗。对经常使用的数据且重要级别较高的数据，归档在高成本、高容量的存储系统上，比如固态硬盘；对经常使用的数据且重要级别一般的数据，归档在低成本、高容量的存储系统上，比如磁盘阵列；对不经常使用但重要级别较高的数据，归档在低成本、高容量的存储系统上，比如磁带设备；对不再使用的数据直接物理删除。最后，制定数据访问策略和安全机制，对具备访问归档数据的用户赋予相关权限。

3.2.7 数据销毁阶段

在数据销毁阶段主要进行数据的销毁和记录，数据销毁阶段是指数据到期后进行销毁的过程，数据生命周期的最后阶段需要安全销毁，需要制定销毁计划，来定义进行数据销毁的时间和方式。通常可以通过机器方式或人工方式进行在线数据销毁和归档数据销毁。同时，为保障后续业务需要，应对销毁的数据内容、销毁时间、销毁方式、销毁人员等信息进行登记，以确保数据销毁的安全性和全面性。

4 数据管理实践与评价

基于科睿唯安核心数据集，从数据接收、数据存储、数据处理、数据计算、数据服务、数据归档、数据销毁 7 个阶段严格按照本文设计的数据管理体系开展基于生命周期的 WOS BP 数据管理实践。然后依照数据管理目标从完整性、唯一性、实时性、有效性、准确性、一致性等 6 个维度进行管理实践与综合评价。最后，依据评价结果得出结论：本文提出的基于生命周期理论的科技文献管理体系适用性良好。下面就具体的评价指标进行说明。

4.1 完整性

完整性是评价数据缺失的情况，包括记录数缺失、字段缺失，属性缺失等，可以在数据接入前或接入后进行监控。以数据字段完整性监测为例，在数据接入后，对 147 个数据项进行监测（图 7），实时评估有值

序号	字段名	描述	有值数据量	空值数据量	有值/空值占比
1	record.article.uid	WOS入藏号	63,567,286	490,397	99.23% 0.77%
2	record.article.sort-date	sort日期	63,567,286	490,397	99.23% 0.77%
3	record.article.cover-date	cover日期	63,567,286	490,397	99.23% 0.77%
4	record.article.pub-year	出版年份	63,567,286	490,397	99.23% 0.77%
5	record.article.pub-month	出版月	52,888,344	11,169,339	82.56% 17.44%
6	record.article.fpage	论文起始页	54,694,828	9,362,855	85.38% 14.62%
7	record.article.lpage	论文结束页	54,694,828	9,362,855	85.38% 14.62%
8	record.article.page-count	论文页数	63,567,286	490,397	99.23% 0.77%
9	record.article.page-range	页码范围	54,694,825	9,362,858	85.38% 14.62%
10	record.article.ea-year	提前出版年	2,246,288	61,811,395	3.51% 96.49%
11	record.article.ea-month	提前出版月	2,246,288	61,811,395	3.51% 96.49%
12	record.article.ea-date	提前出版日期	2,246,288	61,811,395	3.51% 96.49%
13	record.article.type	论文类型	63,567,286	490,397	99.23% 0.77%
14	record.article.type-norm	规范化论文类型	63,567,286	490,397	99.23% 0.77%
15	record.article.abstract	论文摘要	40,405,197	23,652,486	63.08% 36.92%
16	record.article.accession-id	论文传递号	63,567,286	490,397	99.23% 0.77%
17	record.article.doi	论文doi	46,954,405	17,103,278	73.30% 26.70%
18	record.article.pmid	PMID号	25,615,856	38,441,827	39.99% 60.01%
19	record.article.language	文献语种	63,567,286	490,397	99.23% 0.77%
20	record.article.language-norm	规范化文献语种	63,567,286	490,397	99.23% 0.77%
21	record.article.title	文献标题	63,567,286	490,397	99.23% 0.77%

图 7 数据完整性评价

Fig.7 Evaluation of data integrity

数据字段和空值数据字段，对比有值/空值占比，得出数据完整性评价。据统计，截止到2021年12月，147个数据项有值占比为59.75%，必备字段（WOS入藏号、出版年份、文献标题、作者名称、WOS分类、发表期刊标题等）有值占比为99.22%。

4.2 唯一性

唯一性是评价数据重复的情况，包括数据实体是否重复、属性是否重复等，可以在数据接入前或接入后进行监控。针对WOS BP数据设计专业数据字典(图8)，定义12类数据模块，覆盖文献、作者、图书、分类、会议、通讯作者、基金项目、作者机构、出版信息、参考文献、作者信息、发表期刊等内容，通过对数据内容进行监控约束，避免出现数据重复的情况。以数据入藏号为例，数据唯一性达到99.23%。

4.3 实时性

实时性是评价数据及时的情况，是评估数据体现特定时间点的真实程度，包括数据从发表到接收的实时性、数据从接入到服务的实时性，可以在数据接入后进行监控。以数据从接入到服务的实时性为例，以接收第23周数据后和WOS官方6.4日数据量对比：1980—2019年历史数据相差很小，个位数到十位数之间；2020年数据相差百位数；2021年数据相差千位数，是数据处理的正常范围，如图9所示。

4.4 有效性

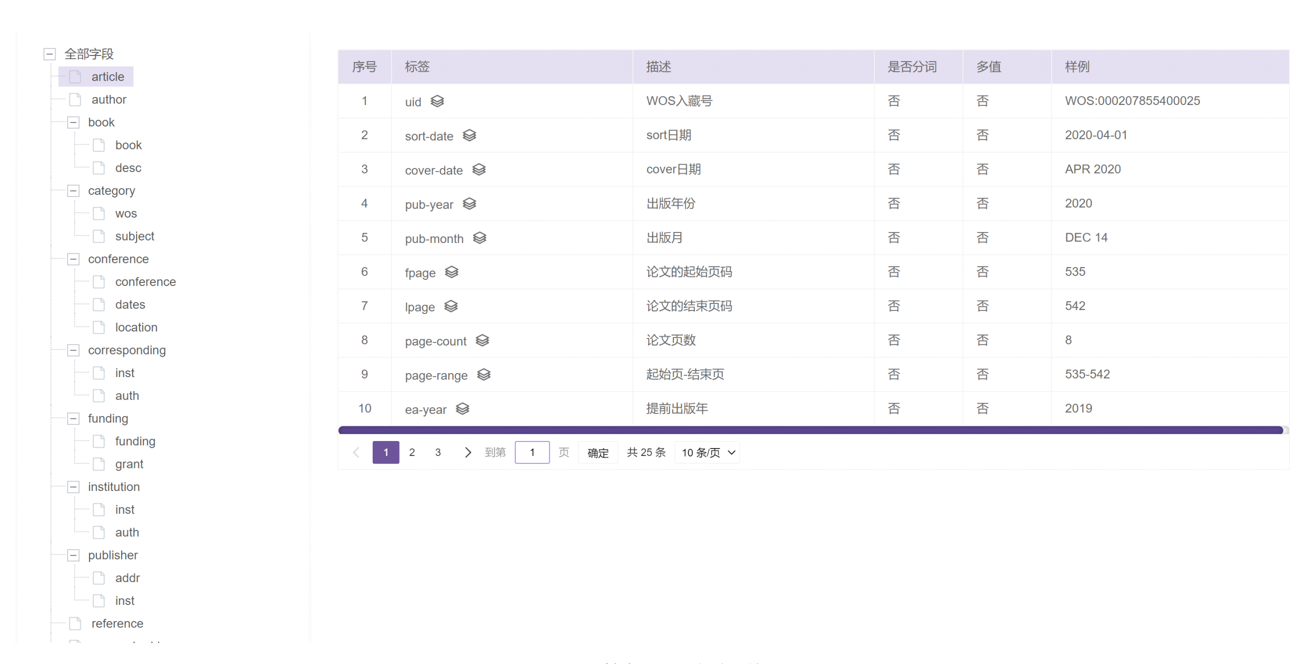
有效性是评价数据项符合规则和定义的情况，包括数据项是否符合类型、格式、种类、范围等约束，是否符合业务逻辑，是否符合值域约束等，可以在数据接入后进行监控。以数据项是否符合类型约束为例，为147个数据项分别定义数据属性区间和类型备选，严格控制每个数据项符合应有的类型约束。

4.5 准确性

准确性是评价数据错误情况，包括数据集合、数据条数、数据项等内容是否与真实数据保持一致，可以在数据接入后进行监控。以数据项准确性评价为例，随机抽取一条数据记录，对比WOS官网数据内容，包括文献信息、发表信息、分类信息、作者信息、基金项目信息等（图10），数据准确性为100%。

4.6 一致性

一致性是评价数据符合标准的情况，也是多次对



序号	标签	描述	是否分词	多值	样例
1	uid	WOS入藏号	否	否	WOS:000207855400025
2	sort-date	sort日期	否	否	2020-04-01
3	cover-date	cover日期	否	否	APR 2020
4	pub-year	出版年份	否	否	2020
5	pub-month	出版月	否	否	DEC 14
6	fpage	论文的起始页码	否	否	535
7	lpage	论文的结束页码	否	否	542
8	page-count	论文页数	否	否	8
9	page-range	起始页-结束页	否	否	535-542
10	ea-year	提前出版年	否	否	2019

图8 数据唯一性评价

Fig.8 Evaluation of data uniqueness

	A	B	C	D	E	F	G	H	Q	R
1	年份	sci	ssci	总量	删除数据	重复数据	有效数据	核对数据	6.4日WOS官网数据	Q-H
2	1980	569397	118874	659291	0	896	658395	658395	658,395	0
3	1981	593282	117733	680126	0	687	679439	679439	679,439	0
4	1982	622978	121864	712594	0	699	711895	711895	711,895	0
5	1983	680727	125181	770534	0	544	769990	769990	769,990	0
6	1984	695453	123652	786234	0	610	785624	785624	785,624	0
7	1985	698076	122363	789210	0	490	788720	788720	788,721	1
8	1986	709704	122323	803285	0	584	802701	802701	802,701	0
9	1987	720625	121463	820106	0	524	819582	819582	819,582	0
10	1988	703499	118695	800862	0	523	800339	800339	800,339	0
11	1989	663400	117641	761246	0	1286	759960	759960	759,960	0
12	1990	692660	119916	788043	0	1853	786190	786190	786,191	1
13	1991	716154	120566	810790	0	4980	805810	805810	805,812	2
14	1992	729255	124172	831990	0	4342	827648	827648	827,649	1
15	1993	770416	126702	872485	0	3469	869016	869016	869,018	2
16	1994	809963	130881	917649	0	3300	914349	914349	914,353	4
17	1995	868253	143220	979829	0	3538	976291	976291	976,291	0
18	1996	912620	148880	1029506	11	3576	1025919	1025919	1,025,924	5
19	1997	947110	145127	1058663	0	3238	1055425	1055425	1,055,425	0
20	1998	955170	142861	1060579	1	3532	1057046	1057046	1,057,047	1
21	1999	985472	145541	1088433	0	3230	1085203	1085203	1,085,205	2
22	2000	995775	153733	1104475	3	3789	1100684	1100684	1,100,684	0
23	2001	992292	144605	1093975	0	3613	1090362	1090362	1,090,363	1
24	2002	1040896	142628	1138935	0	3929	1135006	1135006	1,135,007	1
25	2003	1087331	146856	1183682	8	4680	1178994	1178994	1,178,998	4
26	2004	1176727	152084	1279398	1	4771	1274626	1274626	1,274,628	2
27	2005	1249030	159008	1350455	46	6278	1344153	1344153	1,344,176	23
28	2006	1307401	167301	1412071	46	5463	1406562	1406562	1,406,559	-3
29	2007	1371752	182203	1484686	5	6280	1478402	1478402	1,478,402	0
30	2008	1460614	215664	1603302	236	42603	1560463	1560463	1,560,463	0
31	2009	1587278	231422	1728874	4	96069	1632803	1632803	1,632,802	-1
32	2010	1653621	250284	1803152	10	118915	1684230	1684230	1,684,230	0
33	2011	1754696	263565	1909387	305	144465	1764673	1764673	1,764,673	0
34	2012	1836305	280281	1998884	65	147724	1851127	1851127	1,851,127	0
35	2013	1940271	286675	2098747	61	156579	1942116	1942116	1,942,117	1
36	2014	1998373	291355	2158152	71	163757	1994340	1994340	1,994,297	-43
37	2015	2068182	309819	2240064	560	183779	2055766	2055766	2,055,765	-1
38	2016	2164246	341904	2351614	252	199490	2151960	2151960	2,151,962	2
39	2017	2259533	369273	2439987	645	226127	2213388	2214087	2,214,096	9
40	2018	2355404	394224	2543038	545	241108	2301495	2302569	2,302,558	-11
41	2019	2554931	430379	2755201	8262	323822	2426346	2528997	2,529,007	10
42	2020	3164900	558798	3433717	19581	833204	2583631	2718654	2,719,228	574
43	2021	1301782	221154	1418922	9725	282117	1129430	1130077	1,127,543	-2,534

图 9　数据实时性评价

Fig.9 Evaluation of data real-time

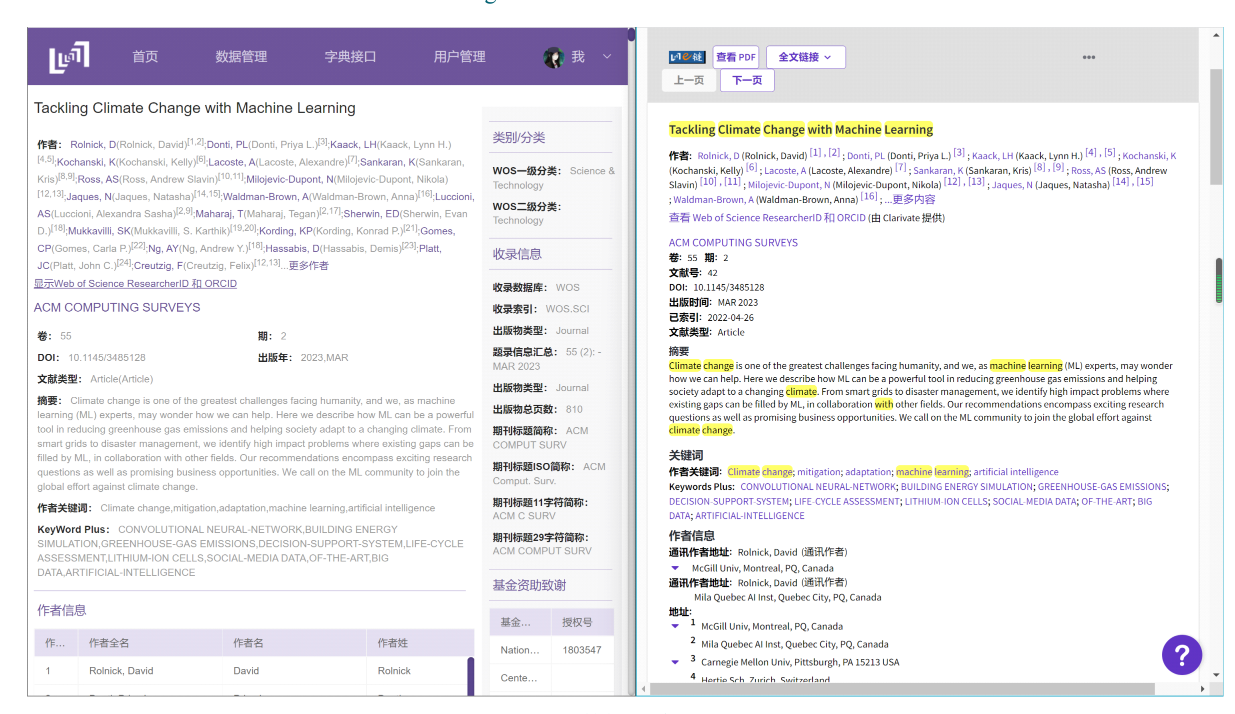


图 10　数据准确性评价

Fig.10 Evaluation of data accuracy

同一数据进行描述而不存在差异的评价，包括数据包是否符合约定形式，数据是否符合数据标准，数据项是否有漏掉或增加等，可以在数据接入前或接入后进行监控。以数据符合标准一致性为例，对接收的 WOS BP 数据进行专项核查，一致性达 90%。

5 总结与展望

本文以数据生命周期为出发点，探究科技文献生命周期管理的关键核心，立足数据管理全流程应用，以科技文献数据为基础，从创建、存储、预处理、计算、服务、归档、销毁 7 个阶段为重点实施步骤进行实践探索，基于 WOS BP 核心数据集实施了上文提出的数据生命周期管理模型，然后从数据质量评估维度进行了完整性、唯一性、实时性、有效性、准确性、一致性等 6 个维度的评价核验，基本解决了科技文献数据在每个生命周期阶段都可以进行有效的管理和应用问题。最终管理模型初具成效，并达到良好的服务效果。

但仍存在很多问题和挑战，在接下来的工作中将进一步完善和改进。首先，在科技文献生命周期管理中集成人工智能技术引擎，紧随国家“新基建”战略部署，让数据管理更智能更全面。其次，在数据生命周期管理中扩展更多种类型和来源的科技文献资源，打通多模态数据智能管理渠道。最后，进一步提升数据生命周期管理的实际应用效果，打造精细化、细粒度的数据形态，提升数据服务水平。

参考文献：

[1] 朱扬勇, 叶雅珍. 从数据的属性看数据资产[J]. 大数据, 2018, 4(6): 65-76.
ZHU Y Y, YE Y Z. Defining data assets based on the attributes of data[J]. Big data research, 2018, 4(6): 65-76.

[2] 云脑数据. 2020 年 25 个令人印象深刻的大数据统计[EB/OL]. [2021-03-26]. https://zhuanlan.zhihu.com/p/360112834.
Dat@mind. 25 impressive big data staticstics in 2020 [EB/OL]. [2021-03-26]. https://zhuanlan.zhihu.com/p/360112834.

[3] 匡华恩. 地质科技文献的管理暨开发利用[J]. 资源信息与工程, 2018, 33(5): 205-206.
KUANG E H. Management and development and utilization of geological scientific and technological documents[J]. Resource information and engineering, 2018, 33(5): 205-206.

[4] Welcome to the cloud security alliance [EB/OL]. [2020-01-21]. https://cloudsecurityalliance.org/.

[5] 刘燕, 杨林, 侯丽, 等. 基于 USGS 生命周期模型的肿瘤流行病学数据管理研究[J]. 中华医学图书情报杂志, 2017, 26(12): 7-14.
LIU Y, YANG L, HOU L, et al. Geographic survey (USGS) data life cycle model-based tumor epidemiology data management[J]. Chinese journal of medical library and information science, 2017, 26(12): 7-14.

[6] USGS data lifecycle overview[EB/OL]. [2018-01-21]. https://www.usgs.gov/data-management/ data-lifecycle.

[7] Document, discover and interoperate[EB/OL]. [2020-01-21]. http://www.icpsr.umich.edu/DDI/index.html

[8] 张迎, 张志平, 梁冰. 科学数据管理应用模式的研究[J]. 情报工程, 2017, 3(4): 71-77.
ZHANG Y, ZHANG Z P, LIANG B. Research on scientific data management application model[J]. Technology intelligence engineering, 2017, 3(4): 71-77.

[9] 国务院办公厅. 国务院办公厅印发《科学数据管理办法》[EB/OL]. [2018-04-02]. http://www.gov.cn/xinwen/2018-04/02/content_5279295.htm.
General Office of the State Council. The General Office of the State Council issued the Measures for the Administration of Scientific Data[EB/OL].[2018-04-02]. http://www.gov.cn/xinwen/2018-04/02/content_5279295.htm.

[10] 国务院办公厅.《国务院办公厅关于印发科学数据管理办法的通知》[EB/OL]. [2018-04-02]. http://www.gov.cn/zhengce/content/2018-04/02/content_5279272.htm.
General Office of the State Council. Notice of the General Office of the State Council on printing and distributing the Measures for the Administration of Scientific Data[EB/OL].[2018-04-02]. http://www.gov.cn/zhengce/content/2018-04/02/content_5279272.htm.

[11] 张洋, 肖燕珠. 生命周期视角下《科学数据管理办法》解读及其启

示[J]. 图书馆学研究, 2019(15): 37–43, 13.

ZHANG Y, XIAO Y Z. Interpretation and englightenment on the Rules of Scientific Data Management from the perspective of life cycle[J]. Research on library science, 2019(15): 37–43, 13.

[12] KUMAR R, ROHITASH K B. Data life cycle management in big data analytics[J]. Procedia computer science, 2020, 173(1): 364–371.

[13] 张培风, 张连分. 全球科研范式变革下的图书馆科学数据管理服务创新——基于数据管理生命周期的视角[J]. 图书馆理论与实践, 2019(5): 39–48.

ZHANG P F, ZHANG L F. The service innovation of library scientific data management under the changes of research paradigm – From the perspective of data life cycle[J]. Library theory and practice, 2019(5): 39–48.

[14] BASHARAT A M, SHEIKH M I, MIR S H. System development life cycle of e–learning content management systems[J]. International journal of knowledge management and practices, 2016, 4(2): 56–66.

[15] 刘南海. 基于 DAMA 体系运营商数据资产管理体系构建研究[J]. 电信网技术, 2016(9): 61–66.

LIU N Y. The research and practice of the data asset management for telecom operator based DAMA[J]. Information and coumunications technology, 2016(9): 61–66.

[16] 张静蓓, 任树怀. 国外科研数据知识库数据质量控制研究[J]. 图书馆杂志, 2016, 35(11): 38–44.

ZHANG J B, REN S H. Studies on data quality control of data repository abroad[J]. Library journal, 2016, 35(11): 38–44.

[17] 常志军, 钱力, 谢靖, 等. 基于分布式技术的科技文献大数据平台的建设研究[J]. 数据分析与知识发现, 2021, 5(3): 69–77.

CHANG Z J, QIAN L, XIE J, et al. Big data platform for sci–rech literature based on distributed technology[J]. Data analysis and knowledge discovery, 2021, 5(3): 69–77.

[18] 张建勇, 于倩倩, 黄永文, 等. NSTL 统一文献元数据标准的设计与思考[J]. 数字图书馆论坛, 2016(2): 33–38.

ZHANG J Y, YU Q Q, HUANG Y W, et al. Metadata standard design of NSTL unified literature[J]. Digital library forum, 2016(2): 33–38.

Scientific and Technical Literature Data Management System Based on Life Cycle Model

CHANG ZhiJun[1,2], XU LiYuan[1*], YU QianQian[1], ZHANG JianYong[1,2], WANG YongJi[3]

(1. National Science Library, Chinese Academy Sciences, Beijing 100190; 2. Department of Library Information and Archives Management, National Science Library, Chinese Academy of Sciences, Beijing 100049; 3. State Key Laboratory of Computer Science Institute of Software, The Chinese Academy of Sciences, Beijing 100190)

Abstract: [Purpose/Significance] Scientific and technical (S&T) literature data resources are characterized with wide coverage, large quantity, many types, fast update and strong timeliness. In order to improve the effect and security of S&T literature data management, this paper studies the S&T literature management system based on the data life cycle model. [Method/Process] This paper explores the management mode of S&T documents, constructs the life cycle system of S&T documents based on the data management process, and expounds the data management tools and methods from the stages of data creation, data storage, data pre-processing, data calculation,

data service, data archiving and data destruction. In the data creation stage, specific data access forms are formulated for different sources and data types, and personalized data creation tools are built to receive data completely. In the data storage stage, a unified document metadata storage system is developed by analyzing the characteristics and shortcomings of various types of data, so as to better explain and organize scientific and technological document data. In the data pre-processing stage, various tools are built to realize the formatting pre-processing, parsing, conversion, structuring and other operations of various types of data. In the data computing stage, data enrichment processing, entity relationship extraction and knowledge graph construction are mainly completed. Data provides services through a unified service interface. Data archiving completes data archiving and saving. In the data destruction phase, unnecessary data is safely destroyed. [Results/Conclusions] In this paper, the management and practice based on the life cycle of S&T literature were first carried out based on the core data set Web Of Science BP data , and then explored from the seven phases of creation, storage, pre-processing, calculation, service, archiving and destruction. Finally, based on the DAMA data quality evaluation principle, the comprehensive evaluation and evaluation of the data management effect were carried out from the six dimensions of integrity, uniqueness, real-time, validity, accuracy and consistency. The receiving integrity of data was 100%, and the non-null integrity of data was 59.75%. The uniqueness of data reached 99.23%. The real time of data was controllable. The validity of data met the constraint conditions. The accuracy of the data reached 100%. The consistency of data reached 90%. It basically solved the problem that data can be effectively managed and applied in each life cycle stage. Finally, the management model was verified to take effect and achieve desirable service effect.

Keywords: life cycle management; scientific and technical (S&T) literature; data management; big data governance; knowledge graph